Positivity-preserving flux limiters for high-order conservative schemes

X. Y. Hu, [a] N. A. Adams [a] and C.-W. Shu [b]

[a] *Lehrstuhl für Aerodynamik und Strömungsmechanik, Technische Universität München, 85748 Garching, Germany*

[b] *Division of Applied Mathematics, Brown University, Providence, RI 02912, USA*

Abstract

In this work a simple method to enforce the positivity-preserving property for general high-order conservative schemes is proposed. The method keeps the original scheme unchanged and detects critical numerical fluxes which may lead to negative density and pressure, and then imposes a cut-off flux limiter to satisfy a sufficient condition for preserving positivity. Though an extra time-step size condition is required to maintain the formal order of accuracy, it is less restrictive than those in previous works. A number of numerical examples suggest that this method, when applied on an essentially non-oscillatory base scheme, can be used to prevent positivity failure when the flow involves vacuum or near vacuum and very strong discontinuities.

Key words: numerical method, compressible flow, high-order conservative scheme, positivity-preserving

1 Introduction

Compressible flow problems are usually solved by conservative schemes. High-order conservative schemes are suitable for simulating flows with both shock waves and rich flow features (acoustic waves, turbulence) since they are capable of handling flow discontinuities and accurately resolve a broad range of length scales. One important issue of high-order conservative schemes is that non-physical negative density or pressure (failure of positivity) can lead to an ill-posed system, which may cause blow-ups of the numerical solution. While for some first-order schemes negative density or pressure can occur when a vacuum or near vacuum is reached, for higher-order conservative schemes positivity failure can also occur due to interpolation errors at or near very strong discontinuities even though the flow physically is far away from vacuum.

It is known that many first order Godunov-type schemes [2, 16, 3] have the so called positivity-preserving property and can maintain positive density and pressure. It has been also proved that some second-order conservative schemes [15, 5] are positivity-preserving with or without a more restrictive Courant-Friedrichs-Lewy (CFL) condition. For even higher-order conservative schemes, Perthame and Shu [11] proved that, given a first-order positivity-preserving scheme, such as Godunov-type schemes, one can always build a higher-order positivity-preserving finite volume scheme under the following constraints: (a) the cell-face values for the numerical flux calculation have positive density and pressure, (b) additional limits on the interpolation under a more restrictive CFL-like condition. With a different interpretation of these constraints based on certain Gauss-Lobatto quadratures, positivity-preserving methods have been successfully developed for high-order discontinuous Galerkin (DG)

2

methods [18] and weighted essentially non-oscillatory (WENO) finite volume and finite difference schemes [19, 20].

In this paper, we propose an alternative method to enforce the positivity-preserving property with a simple cut-off flux limiter. The flux limiter first detects critical numerical fluxes which may lead to negative density and pressure, then limits these fluxes to satisfy a sufficient condition for preserving positivity. Unlike the approaches in [18, 19, 20], in which positivity-preserving and the maintenance of high order accuracy are considered simultaneously when designing the limiter, here we design the cut-off flux limiter to satisfy positivity only, and then prove *a posteriori* the maintenance of high order accuracy under a time step restriction. It appears that, in our numerical experiments, a much less restrictive time-step size condition is sufficient for preserving positivity without destroying overall accuracy. An advantage of the approach in this paper is that the cut-off limiter is directly applied to the numerical flux and it can be applied to arbitrary high-order conservative schemes.

2   Method

For presentation of the positivity-preserving flux limiters we assume that the fluid is inviscid and compressible, described by the one-dimensional Euler equations as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = 0, \tag{1}$$

where $\mathbf{U} = (\rho, m, E)^T$, and $\mathbf{F}(\mathbf{U}) = [m, \rho u^2 + p, (E + p)u]^T$. This set of equations describes the conservation laws for mass density ρ, momentum density

3

$m \equiv \rho u$ and total energy density $E = \rho e + \rho u^2/2$, where e is the internal energy per unit mass. To close this set of equations, the ideal-gas equation of state $p = (\gamma - 1)\rho e$ with a constant γ is used. Note that the density and pressure have the relations with the conservative variables as

$$\rho(\mathbf{U}) = \rho, \quad p(\mathbf{U}) = (\gamma - 1)\left(E - \frac{1}{2}\frac{m^2}{\rho}\right). \tag{2}$$

It is easy to find that they are locally Lipschitz continuous, i.e.

$$|\rho(\mathbf{U}_2) - \rho(\mathbf{U}_1)| \leq L_\rho ||\mathbf{U}_2 - \mathbf{U}_1||, \tag{3}$$
$$|p(\mathbf{U}_2) - p(\mathbf{U}_1)| \leq L_p ||\mathbf{U}_2 - \mathbf{U}_1||, \text{ if } \rho(\mathbf{U}_1) > 0, \ \rho(\mathbf{U}_2) > 0, \tag{4}$$

where L_ρ and L_p are Lipschitz constants. For $1 \geq \theta \geq 0$, $\rho(\mathbf{U})$ and $p(\mathbf{U})$ have the properties

$$\rho\left[(1 - \theta)\mathbf{U}_1 + \theta\mathbf{U}_2\right] = (1 - \theta)\rho(\mathbf{U}_1) + \theta\rho(\mathbf{U}_2), \tag{5}$$
$$p\left[(1 - \theta)\mathbf{U}_1 + \theta\mathbf{U}_2\right] \geq (1 - \theta)p(\mathbf{U}_1) + \theta p(\mathbf{U}_2), \text{ if } \rho(\mathbf{U}_1), \ \rho(\mathbf{U}_2) > 0, \tag{6}$$

where Eq. (5) is straightforward and Eq. (6) is implied by the Jensen's inequality since $p(\mathbf{U})$ is a concave function.

2.1 Finite-volume and finite-difference conservative schemes

When Eq. (1) is discretized within the spatial domain such that $x_i = i\Delta x$, $i = 0, ..., N$, where Δx is the spatial step, a general explicit kth-order conservative scheme with Euler-forward time integration can be written as

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n + \lambda\left(\hat{\mathbf{F}}_{i-1/2} - \hat{\mathbf{F}}_{i+1/2}\right), \tag{7}$$

where the superscript n and $n + 1$ represent the old and new time steps, respectively, and $\lambda = \Delta t/\Delta x$, where Δt is time-step size. Note that with the

4

CFL condition

$$\Delta t = \frac{\text{CFL} \cdot \Delta x}{(|u| + c)_{\max}}, \tag{8}$$

where $c = \sqrt{\gamma p/\rho}$ is the sound speed and the CFL number $0 < \text{CFL} < 1$, one has the relation

$$\lambda = \frac{\text{CFL}}{(|u| + c)_{\max}}. \tag{9}$$

For a finite-volume scheme, $\mathbf{U}_i^n$ and $\mathbf{U}_i^{n+1}$ are the cell averaged conservative variables on the cell i defined on the computational cell between $(i - 1/2)\Delta x$ and $(i + 1/2)\Delta x$, i.e. $I_i = [x_{i-1/2}, x_{i+1/2}]$, $\hat{\mathbf{F}}_{i\pm1/2} = \mathbf{F}_{i\pm1/2} + \mathbf{O}(\Delta x^{k+1})$ are the numerical fluxes, which are based on the cell-face values $\mathbf{U}_{i\pm1/2}$ reconstructed from the cell averages $\{\mathbf{U}_j\}$ and $\mathbf{F}_{i\pm1/2} = \mathbf{F}(\mathbf{U}_{i\pm1/2})$.

For a finite-difference scheme, $\mathbf{U}_i^n$ and $\mathbf{U}_i^{n+1}$ are the nodal values, and $(\hat{\mathbf{F}}_{i+1/2} - \hat{\mathbf{F}}_{i-1/2})/\Delta x$ is a kth order approximation to $\partial \mathbf{F}(\mathbf{U})/\partial x$ at $x = x_i$. Assume there exists a function $\mathbf{H}(x)$ depending on Δx such that

$$\mathbf{F}\left[\mathbf{U}(x)\right] = \frac{1}{\Delta x} \int\limits_{x-\Delta x/2}^{x+\Delta x/2} \mathbf{H}(\xi)d\xi, \tag{10}$$

then the same reconstruction procedure in a finite-volume scheme can be used to obtain the numerical fluxes $\hat{\mathbf{F}}_{i\pm1/2} = \mathbf{H}_{i\pm1/2} + O(\Delta x^{k+1})$ based on the cell-face values of $\mathbf{H}(x)$ reconstructed from its cell-average values $\mathbf{F}\left[\mathbf{U}_j\right] = \int_{x_j-\Delta x/2}^{x_j+\Delta x/2} \mathbf{H}(\xi)d\xi/\Delta x$. We refer to [14] for the discussion of this formulation of conservative finite difference schemes.

2.2 Positivity preserving cut-off flux limiter

The positivity-preserving property for the scheme Eq. (7) refers to the property that the density and pressure are positive for $\mathbf{U}_i^{n+1}$ when $\mathbf{U}_i^n$ has positive density and pressure. Since Eq. (7) can be rewritten as a convex combination

$$
\begin{aligned}
\mathbf{U}_i^{n+1} &= \frac{1}{2}\left(\mathbf{U}_i^n + 2\lambda\hat{\mathbf{F}}_{i-1/2}\right) + \frac{1}{2}\left(\mathbf{U}_i^n - 2\lambda\hat{\mathbf{F}}_{i+1/2}\right) \\
&= \frac{1}{2}\mathbf{U}_i^- + \frac{1}{2}\mathbf{U}_i^+,
\end{aligned} \tag{11}
$$

a sufficient condition for preserving positivity is that $\mathbf{U}_i^\pm$ have positive density and pressure, i.e. $g(\mathbf{U}_i^\pm) > 0$, where g represents ρ and p. Since the first-order Lax-Friedrichs flux

$$
\hat{\mathbf{F}}_{i+1/2}^{LF} = \frac{1}{2}\left[\mathbf{F}_i + \mathbf{F}_{i+1} + (|u| + c)_{max}(\mathbf{U}_i^n - \mathbf{U}_{i+1}^n)\right] \tag{12}
$$

has the property $g(\mathbf{U}_i^{LF,\pm}) = g(\mathbf{U}_i^n \mp 2\lambda\hat{\mathbf{F}}_{i\pm1/2}^{LF}) > 0$, under an additional CFL condition

$$
\text{CFL} \leq \frac{1}{2} \tag{13}
$$

(see [18]), a straightforward way to ensure positivity is to limit the magnitude of $\hat{\mathbf{F}}_{i+1/2}$ by utilizing the properties in Eqs. (5) and (6). The positive density is first enforced by:

Cut-off flux limiter for positive density

1. For all i: initialize $\theta_{i+1/2}^+ = 1$, $\theta_{i+1/2}^- = 1$.

2. If $\rho(\mathbf{U}_i^+) < \epsilon_\rho$, solve $\theta_{i+1/2}^+$ from $(1 - \theta_{i+1/2}^+)\rho(\mathbf{U}_i^{LF,+}) + \theta_{i+1/2}^+\rho(\mathbf{U}_i^+) = \epsilon_\rho$.

3. If $\rho(\mathbf{U}_{i+1}^-) < \epsilon_\rho$, solve $\theta_{i+1/2}^-$ from $(1 - \theta_{i+1/2}^-)\rho(\mathbf{U}_{i+1}^{LF,-}) + \theta_{i+1/2}^-\rho(\mathbf{U}_{i+1}^-) = \epsilon_\rho$.

4. Set $\theta_{\rho,i+1/2} = \min(\theta_{i+1/2}^+, \theta_{i+1/2}^-)$, $\hat{\mathbf{F}}_{i+1/2}^* = (1 - \theta_{\rho,i+1/2})\hat{\mathbf{F}}_{i+1/2}^{LF} + \theta_{\rho,i+1/2}\hat{\mathbf{F}}_{i+1/2}$.

6

Here, $\epsilon_\rho = \min\{10^{-13}, \rho^0_{min}\}$, where ρ^0_{min} is the minimum density in the initial condition, $\hat{\mathbf{F}}^*_{i+1/2}$ is the limited flux, $0 \le \theta^\pm_{i+1/2} \le 1$ are the limiting factors corresponding to the two neighboring cells, which share the same flux $\hat{\mathbf{F}}_{i+1/2}$. After applying this flux limiter, Eq. (11) becomes

$$
\begin{aligned}
\mathbf{U}^{n+1}_i &= \frac{1}{2}\left(\mathbf{U}^n_i + 2\lambda\hat{\mathbf{F}}^*_{i-1/2}\right) + \frac{1}{2}\left(\mathbf{U}^n_i - 2\lambda\hat{\mathbf{F}}^*_{i+1/2}\right) \\
&= \frac{1}{2}\mathbf{U}^{*,-}_i + \frac{1}{2}\mathbf{U}^{*,+}_i.
\end{aligned}
\tag{14}
$$

Clearly, by Eq. (5), both $\mathbf{U}^{*,-}_i$ and $\mathbf{U}^{*,+}_i$ have positive density, so does $\mathbf{U}^{n+1}_i$. The positive pressure is further enforced by:

Cut-off flux limiter for positive pressure

1. For all i: initialize $\theta^+_{i+1/2} = 1$, $\theta^-_{i+1/2} = 1$.
2. If $p(\mathbf{U}^{*,+}_i) < \epsilon_p$, solve $\theta^+_{i+1/2}$ from $(1-\theta^+_{i+1/2})p(\mathbf{U}^{LF,+}_i)+\theta^+_{i+1/2}p(\mathbf{U}^{*,+}_i) = \epsilon_p$.
3. If $p(\mathbf{U}^{*,-}_{i+1}) < \epsilon_p$, solve $\theta^-_{i+1/2}$ from $(1-\theta^-_{i+1/2})p(\mathbf{U}^{LF,-}_{i+1})+\theta^-_{i+1/2}p(\mathbf{U}^{*,-}_{i+1}) = \epsilon_p$.
4. Set $\theta_{p,i+1/2} = \min(\theta^+_{i+1/2}, \theta^-_{i+1/2})$, $\hat{\mathbf{F}}^{**}_{i+1/2} = (1-\theta_{p,i+1/2})\hat{\mathbf{F}}^{LF}_{i+1/2}+\theta_{p,i+1/2}\hat{\mathbf{F}}^*_{i+1/2}$.

Again, $\epsilon_p = \min\{10^{-13}, p^0_{min}\}$, where p^0_{min} is the minimum pressure in the initial condition, and $\hat{\mathbf{F}}^{**}_{i+1/2}$ is the further limited flux. After applying this flux limiter, Eq. (14) becomes

$$
\begin{aligned}
\mathbf{U}^{n+1}_i &= \frac{1}{2}\left(\mathbf{U}^n_i + 2\lambda\hat{\mathbf{F}}^{**}_{i-1/2}\right) + \frac{1}{2}\left(\mathbf{U}^n_i - 2\lambda\hat{\mathbf{F}}^{**}_{i+1/2}\right) \\
&= \frac{1}{2}\left(\mathbf{U}^{**,-}_i + \mathbf{U}^{**,+}_i\right).
\end{aligned}
\tag{15}
$$

Clearly, by Eqs. (5) and (6), both $\mathbf{U}^{**,-}_i$ and $\mathbf{U}^{**,+}_i$ have positive density and pressure, so does $\mathbf{U}^{n+1}_i$. Note that these limiters can be applied at each sub-stage of a TVD Runge-Kutta [13] method, which is a convex combination of Euler-forward time steps.

2.3 Consistency and accuracy

Now we address two important issues for the cut-off flux limiter. First, the limited flux is a consistent flux since it is the convex combination of two consistent fluxes, i.e. the first-order Lax-Friedrichs flux $\mathbf{U}_{i+1/2}^{LF}$ and the original high-order numerical flux $\hat{\mathbf{F}}_{i+1/2}^{o}$, which represents $\hat{\mathbf{F}}_{i+1/2}$ and $\hat{\mathbf{F}}_{i+1/2}^{*}$. Second, when the limiter is active, the difference between the original flux $\hat{\mathbf{F}}_{i+1/2}^{o}$ and the limited flux $\hat{\mathbf{F}}_{i+1/2}^{lim}$ representing $\hat{\mathbf{F}}_{i+1/2}^{*}$ and $\hat{\mathbf{F}}_{i+1/2}^{**}$ is

$$||\hat{\mathbf{F}}_{i+1/2}^{lim} - \hat{\mathbf{F}}_{i+1/2}^{o}|| = (1 - \theta_{g,i+1/2})||\hat{\mathbf{F}}_{i+1/2}^{o} - \hat{\mathbf{F}}_{i+1/2}^{LF}||. \tag{16}$$

We only need to consider accuracy maintenance when $\theta_{g,i+1/2} < 1$, for otherwise the limiter does not take any effect. Without loss of generality we may assume $\theta_{g,i+1/2} = \theta_{g,i+1/2}^{+}$. In this situation we have $g(\mathbf{U}_{i}^{o,+}) < \epsilon_g$, in which $\mathbf{U}_{i}^{o,+}$ represents $\mathbf{U}_{i}^{+}$ and $\mathbf{U}_{i}^{*,+}$, and ϵ_g is negligibly small, and

$$1 - \theta_{g,i+1/2} = \frac{\epsilon_g - g(\mathbf{U}_{i}^{o,+})}{g(\mathbf{U}_{i}^{LF,+}) - g(\mathbf{U}_{i}^{o,+})} \approx \frac{-g(\mathbf{U}_{i}^{o,+})}{g(\mathbf{U}_{i}^{LF,+}) - g(\mathbf{U}_{i}^{o,+})} \leq \frac{|g(\mathbf{U}_{i}^{o,+})|}{g(\mathbf{U}_{i}^{LF,+})}. \tag{17}$$

Since $\hat{\mathbf{F}}_{i+1/2}^{o}$ and $\hat{\mathbf{F}}_{i+1/2}^{LF}$ are both bounded in smooth regions, it is sufficient to show that the accuracy is not destroyed if the limiting factor satisfies

$$1 - \theta_{g,i+1/2} = O(\Delta x^{k+1}), \tag{18}$$

a sufficient condition for which would be $|g(\mathbf{U}_{i}^{o,+})| = O(\Delta x^{k+1})$ and $g(\mathbf{U}_{i}^{LF,+})$ is bounded away from zero.

Similar to Zhang and Shu [18], we assume the exact solution $\mathbf{U}(x)$ is smooth and $g(\widetilde{\mathbf{U}}_i) \geq M$, where $\widetilde{\mathbf{U}}_i$ is either the cell-average (for the finite-volume scheme) or the nodal value (for the finite-different scheme) of the exact solution

$\mathbf{U}(x)$ and $M > 0$ is a constant. Since $g(\mathbf{U}_i)$ is obtained from a kth order approximation, one has $g(\mathbf{U}_i) \geq M - O(\Delta x^{k+1}) > M/2$ if Δx is sufficiently small, therefore

$$
\begin{aligned}
g(\mathbf{U}_i^{LF,+}) &= g\left[(1 - \hat{w})\mathbf{U}_i + \hat{w}\left(\mathbf{U}_i - \frac{2\lambda}{\hat{w}}\hat{\mathbf{F}}_{i+1/2}^{LF}\right)\right] \\
&\geq (1 - \hat{w})g(\mathbf{U}_i) + \hat{w}g\left(\mathbf{U}_i - \frac{2\lambda}{\hat{w}}\hat{\mathbf{F}}_{i+1/2}^{LF}\right) \\
&\geq \frac{(1 - \hat{w})}{2}M > 0,
\end{aligned}
\tag{19}
$$

where $1 > \hat{w} > 0$ is a constant, under an extra CFL condition

$$
\text{CFL} \leq \frac{\hat{w}}{2}.
\tag{20}
$$

Furthermore, one has

$$
\begin{aligned}
\mathbf{U}_i^{o,+} &= \mathbf{U}_i^n - 2\lambda\hat{\mathbf{F}}_{i+1/2}^o \\
&= \mathbf{U}_i^{LF,+} + 2\lambda\left(\hat{\mathbf{F}}_{i+1/2}^{LF} - \hat{\mathbf{F}}_{i+1/2}^o\right) \\
&= \mathbf{U}_i^{LF,+} + 2\lambda\left(\hat{\mathbf{F}}_{i+1/2}^{LF} - \widetilde{\mathbf{F}}_{i+1/2}\right) + \mathbf{O}(\Delta x^{k+1}),
\end{aligned}
\tag{21}
$$

where $\widetilde{\mathbf{F}}_{i+1/2} = \mathbf{F}_{i+1/2}$ for the finite-volume scheme, and $\widetilde{\mathbf{F}}_{i+1/2} = \mathbf{H}_{i+1/2}$ for the finite-difference scheme. Let $\mathbf{U}_i^s = \mathbf{U}_i^{LF,+} + 2\lambda\left(\hat{\mathbf{F}}_{i+1/2}^{LF} - \widetilde{\mathbf{F}}_{i+1/2}\right)$, and with Eqs. (3) and (4), one has

$$
|g(\mathbf{U}_i^s) - g(\mathbf{U}_i^{o,+})| \leq L_g\|\mathbf{U}_i^{o,+} - \mathbf{U}_i^s\| = O(\Delta x^{k+1}),
\tag{22}
$$

where L_g is the Lipschitz constant. Note that the first term of $\mathbf{U}_i^s$ has positive density and pressure. For the second term, since the first-order Lax-Friedrichs flux $\hat{\mathbf{F}}_{i+1/2}^{LF}$ is a first order approximation to the exact flux $\widetilde{\mathbf{F}}_{i+1/2}$, that is $\|\hat{\mathbf{F}}_{i+1/2}^{LF} - \widetilde{\mathbf{F}}_{i+1/2}\| = O(\Delta x)$. With bounded $g(\mathbf{U}_i^{LF,+})$ from Eq. (19), one has

$$
\rho(\mathbf{U}_i^s) \geq \frac{(1 - \hat{w})}{2}M - O(\Delta x) \geq \frac{(1 - \hat{w})}{4}M > 0
$$

for sufficiently small Δx, according to Eq. (3), and furthermore $p(\mathbf{U}_i^s) > \epsilon_p$ according to Eq. (4). Since $g(\mathbf{U}_i^s) > \epsilon_g$ and $g(\mathbf{U}_i^{o,+}) < \epsilon_g$, i.e. $\rho(\mathbf{U}_i^+) < \epsilon_\rho$ while enforcing positive density and $p(\mathbf{U}_i^{*,+}) < \epsilon_p$ but $\rho(\mathbf{U}_i^{*,+}) > \epsilon_\rho$ while enforcing positive pressure, Eq. (22) leads to $|g(\mathbf{U}_i^{o,+})| = O(\Delta x^{k+1})$. Hence, we have proved that the cut-off flux limiter preserves high-order accuracy.

Note that, for given values of M and grid size, Eqs. (19) and (21) suggest that the errors introduced by the the cut-off flux limiter decrease with the time-step sizes. Also note that, the condition Eq. (20) is less restrictive than the time-step size conditions in Refs. [18, 19, 20], and is desirable for higher computational efficiency.

2.4 Assessment of accuracy

As a simple way to test the accuracy of the present flux limiters, we consider the one-dimensional linear advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0 \tag{23}$$

with initial condition $u(x) > 0$. Applying the cut-off flux limiter to preserve positivity results in the limiter (denoted as HAS)

$$f_{i+1/2}^* = \theta(u_{i+1/2}^- - u_i^n) + u_i^n, \quad \theta = \min\left\{\frac{u_i^n}{u_i^n - u_{\min}}, 1\right\},$$
$$u_{\min} = \min\left\{u_i^n - 2\lambda u_{i+1/2}^-, u_{i+1}^n + 2\lambda u_{i+1/2}^-, 10^{-13}\right\}. \tag{24}$$

Here $u_{i+1/2}^-$ is the approximated upwind flux at the cell face $i + 1/2$. Note that only one of $u_i^n - 2\lambda u_{i+1/2}^-$ or $u_{i+1}^n + 2\lambda u_{i+1/2}^-$ being negative will activate the limiter. The limiter of Zhang and Shu [19] (denoted as ZS) for Eq. (23) can

be written as

$$f^*_{i+1/2} = \theta(u^-_{i+1/2} - u^n_i) + u^n_i, \quad \theta = \min\left\{\frac{u^n_i}{u^n_i - u_{\min}}, 1\right\},$$

$$u_{\min} = \min\left\{\frac{u^n_i - \hat{w}_1(u^+_{i-1/2} + u^-_{i+1/2})}{1 - 2\hat{w}_1}, u^+_{i-1/2}, u^-_{i+1/2}, 10^{-13}\right\}, \quad (25)$$

where $u^+_{i-1/2}$ and $u^-_{i+1/2}$ are the high-order approximations of the cell-face values at $u(x_{i-1/2})$ and $u(x_{i+1/2})$ within cell i. For Eq. (23), one has $u^+_{i-1/2} = u_{i+1/2}$. Comparing $u_{\min}$ in Eqs. (24) and (25), it can be observed that the HAS limiter does not directly constrain the cell-face values to be non-negative.

To further illustrate the accuracy of the HAS limiter and its relation to the ZS limiter, we compute the advection of a function $u = 1 + 10^{-6} + \cos(2\pi x)$ in domain $[0, 1]$ with a fifth-order conservative finite difference WENO-5 scheme [7] with third-order TVD Runge-Kutta time integration [13]. A periodic boundary condition is applied at $x = 0$ and $x = 1$. The final time is $t = 1$, which corresponds to one period. This problem is computed on different grids with $N = 50$, 100, 200, 400 and 800 grid points. Figure 1a shows the error distributions for the results on 200 grid points. It can be observed that if the maximum admissible CFL number of 0.5 is used, the HAS limiter produces larger errors than the ZS limiter. However, the HAS limiter is already as accurate as the ZS limiter when a smaller CFL number of 1/12, which corresponds to the maximum admissible value for the latter, is used. If the time-step size is decreased further, errors produced by the ZS limiter do not change considerably, whereas the errors produced by the HAS limiter decrease further. This behavior is also shown in Fig. 1b for the evolution of the L_∞ error with decreasing grid size. Here, the time-step size $\Delta t = 0.5\Delta x^{5/3}$ is used to keep the spatial errors dominant. Note that Fig. 1b clearly shows that the theoretical

11

order of accuracy is achieved.

2.5 Extension to multiple dimensions

To present the extension of the positivity-preserving flux limiters to multiple dimensions we consider the two-dimensional Euler equation

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial y} = 0. \tag{26}$$

where $\mathbf{U} = (\rho, \rho u, \rho v, E)^T$, $\mathbf{F}(\mathbf{U}) = [\rho u, \rho u^2 + p, \rho u v, (E + p)u]^T$ and $\mathbf{G}(\mathbf{U}) = [\rho v, \rho v u, \rho v^2 + p, (E+p)v]^T$. Compared to the one-dimensional equation Eq. (1), the momentum density is $\rho \mathbf{v} = (\rho u, \rho v)$, where u and v are velocities in the x and y directions, respectively, and the total energy density is $E = \rho e + \rho \mathbf{v}^2/2$. As an extension of Eq. (11), the conservative scheme for Eq. (26) can be rewritten as a convex combination

$$\begin{aligned}
\mathbf{U}_{i,j}^{n+1} = {} & \frac{\alpha_x}{2} \left(\mathbf{U}_{i,j}^n + 2\lambda_x \hat{\mathbf{F}}_{i-1/2,j} \right) + \frac{\alpha_x}{2} \left(\mathbf{U}_{i,j}^n - 2\lambda_x \hat{\mathbf{F}}_{i+1/2,j} \right) \\
& + \frac{\alpha_y}{2} \left(\mathbf{U}_{i,j}^n + 2\lambda_y \hat{\mathbf{G}}_{i,j-1/2} \right) + \frac{\alpha_y}{2} \left(\mathbf{U}_{i,j}^n - 2\lambda_y \hat{\mathbf{G}}_{i,j+1/2} \right),
\end{aligned} \tag{27}$$

where $\lambda_x = \Delta t/\Delta x \alpha_x$ and $\lambda_y = \Delta t/\Delta y \alpha_y$, $\alpha_x + \alpha_y = 1$, with $\alpha_x > 0$ and $\alpha_y > 0$ being partitions of the contribution in the x and y directions. A simple way to obtain this partition is to set $\alpha_x = \alpha_y = 1/2$ as in Zhang and Shu [18, 20]. Another straightforward way to determine α_x and $\alpha_y = 1 - \alpha_x$ is

$$\alpha_x = \frac{\tau_x}{\tau_x + \tau_y}, \quad \tau_x = \frac{(|u| + c)_{\max}}{\Delta x}, \quad \tau_y = \frac{(|v| + c)_{\max}}{\Delta y}. \tag{28}$$

Note that, since the time-step size for integrating Eq. (27) is given by

$$\Delta t = \frac{\text{CFL}}{\tau_x + \tau_y}, \tag{29}$$

one has the relation

$$\lambda_x = \frac{\text{CFL}}{(|u| + c)_{\text{max}}} \quad \text{and} \quad \lambda_y = \frac{\text{CFL}}{(|v| + c)_{\text{max}}}, \tag{30}$$

which gives an extended form from Eq. (9). Also note that, since the components in Eq. (27) and Eq. (11) have the same form, it is straightforward to implement the positivity-preserving flux limiters in a dimension-by-dimension fashion.

3   Test cases

In the following, we illustrate that a number of typical numerical test cases, where the original high-order conservative schemes fail, can be simulated by using the proposed positivity-preserving flux limiters. For the first type of cases involving vacuum or near vacuum, the flux limiters are combined with the finite difference WENO-5 scheme [7], which is a shock-capturing scheme with fifth-order accuracy for smooth solutions. For the second type of cases involving very strong discontinuities, the flux limiters are combined with the WENO-CU6-M1 scheme [4], which can be used for implicit large eddy simulation (LES) of turbulent flow and has sixth-order accuracy for smooth solutions. For both variants of the WENO schemes the Roe approximation is used for the characteristic decomposition at the cell faces, the Lax-Friedrichs formulation is used for the numerical fluxes, and the third-order TVD Runge-Kutta scheme is used for time integration [13]. If not mentioned otherwise, the computations are carried out with a CFL number of 0.5.

3.1 One-dimensional problems involving vacuum or near vacuum

Here we show that the proposed method passes two one-dimensional test problems involving vacuum or near vacuum: the double rarefaction problem [5], where a vacuum occurs, and the planar Sedov blast-wave problem [12, 20], where a point-blast wave propagates. For the first problem, the initial condition is

$$(\rho, u, p) = \begin{cases} (1, -2, 0.1) & \text{if } 0 < x < 0.5 \\ (1, 2, 0.1) & \text{if } 1 > x > 0.5 \end{cases},$$

$\Delta x = 2.5 \times 10^{-3}$ and the final time is $t = 0.1$. For the second problem, the initial condition is

$$(\rho, u, p) = \begin{cases} (1, 0, 4 \times 10^{-13}) & \text{if } 0 < x < 2 - 0.5\Delta x, \quad 2 + 0.5\Delta x < x < 4 \\ (1, 0, 2.56 \times 10^{8}) & \text{if } 2 - 0.5\Delta x < x < 2 + 0.5\Delta x \end{cases},$$

$\Delta x = 5 \times 10^{-3}$ and the final time is $t = 10^{-3}$.

Figure 2 gives the computed pressure, density and velocity distributions, which show good agreement with the exact solutions. Although a vacuum occurs in the solution of the double rarefaction problem, the results still exhibit accurate density and pressure profiles in the rarefaction-wave regions. As a vacuum occurs, the solution at the center of the domain strictly speaking has no physical meaning. Note that compared to Zhang and Shu [20] (see their Fig. 5.1 (right)) for the planar Sedov blast-wave problem a slightly sharper blast wave is obtained in the present results. This may be due to the fact that Zhang and Shu [20] have modified the original Lax-Friedrichs flux to use a single maximum signal speed other than the respective maximum eigenvalues.

3.2 Two-dimensional problems involving vacuum or near vacuum

We consider two two-dimensional problems involving vacuum or near vacuum. The first problem is the two-dimensional Sedov problem which has been studied in Zhang and Shu [18, 20]. The computation is performed on the domain $[0,0] \times [1.1, 1.1]$, where a high pressure region occupies the computation cell at the lower-left corner. The initial condition is given by

$$(\rho, u, v, p) = \begin{cases} (1,0,0,4 \times 10^{-13}) & \text{if } x > \Delta x, \quad y > \Delta y \\ (1,0,0, \frac{9.79264}{\Delta x \Delta y} \times 10^4) & \text{else} \end{cases},$$

where $\Delta x = \Delta y = 1.1/160$. The final time is $t = 1.0 \times 10^{-3}$. A reflective boundary condition is applied at the lower and left boundaries, and an outflow condition is applied at the right and upper boundaries. Figure 3 gives the computed density profiles. One can observe that these results are in very good agreement with the exact solution.

The second problem is the Mach-2000 jet problem, which has been computed in Zhang and Shu [18, 19, 20]. The computation is performed on the domain $[0,1] \times [0,0.25]$, Initially, the entire domain is filled with ambient gas with $(\rho, u, v, p) = (0.5, 0, 0, 0.4127)$. A reflective condition is applied at the lower boundary, an outflow condition is applied at the right and upper boundaries, and an inflow condition is applied at the left boundary with states $(\rho, u, v, p) = (5, 800, 0, 0.4127)$ if $y < 0.05$ and $(\rho, u, v, p) = (0.5, 0, 0, 0.4127)$ otherwise. A CFL number of 0.25 is used and the final time is 0.001. Since $\gamma = 5/3$ is used, the speed of the jet 800 gives about Mach 2100 with respect to the sound speed in the jet gas. Figure 4 gives the computed density and pressure profiles in logarithmic scale. One can observe that these results are in very good agreement with those in Zhang and Shu [18] (their Fig. 4.6) computed

with the same resolution.

3.3 One-dimensional problems involving very strong discontinuities

We show that, combined with the proposed flux limiters, the WENO-CU6-M1 scheme passes two one-dimensional test problems, which cannot be computed with the original scheme without limiting, involving very strong discontinuities: the two blast-wave interaction problem [17], and the Le Blanc problem [10]. The latter is an extreme shock-tube problem. For the first problem, the initial condition is

$$(\rho, u, p) = \begin{cases} (1, 0, 1000) & \text{if } 0 < x < 0.1 \\ (1, 0, 0.01) & \text{if } 0.1 < x < 0.9 \\ (1, 0, 100) & \text{if } 1 > x > 0.9 \end{cases},$$

$\Delta x = 2.5 \times 10^{-3}$, and the final time is $t = 0.038$. Reflective boundary conditions are applied at both $x = 0$ and $x = 1$. The reference "exact" solution is a high-resolution numerical solution on 3200 grid points calculated by the WENO-CU6 scheme [6]. For the second problem, the initial condition is

$$(\rho, u, p) = \begin{cases} (1, 0, \frac{2}{3} \times 10^{-1}) & \text{if } 0 < x < 3 \\ (10^{-3}, 0, \frac{2}{3} \times 10^{-10}) & \text{if } 3 < x < 9 \end{cases},$$

$\gamma = 5/3$, $\Delta x = 9/800$ and the final time is $t = 6$.

Figure 5 gives the computed pressure, density and velocity distributions, although at relatively low resolution, which show a good agreement with the exact or reference solutions. The magnitudes of the small over-shoots (see Fig. 5(left)) and the small errors at the shock position (see Fig. 5(right)) decrease when the resolution is increased (not shown here). For the two blast-wave interaction problem the present results are comparable to those obtained by

16

the WENO-CU6-M2 scheme [4] at the same resolution (see their Fig. 3). Note that the WENO-CU6-M2 scheme stabilizes for very strong discontinuities in a different way, but still cannot compute the Le Blanc problem.

3.4   Two-dimensional problems involving very strong discontinuities

We first consider the problem from Woodward and Colella [17] on the double Mach reflection of a strong shock. A Mach 10 shock in air is reflected from the wall with incidence angle of 60°. The initial condition is

$$(\rho, u, v, p) = \begin{cases} (1.4, 0, 0, 1) & \text{if } y < 1.732(x - 0.1667) \\ (8, 7.145, -4.125, 116.8333) & \text{else} \end{cases},$$

and the final time is $t = 0.2$. The computational domain for this problem is $[0, 0] \times [4, 1]$. Initially, the shock extends from the point $x = 0.1667$ at the bottom to the top of the computational domain. Along the bottom boundary, at $y = 0$, from $x = 0$ to $x = 0.1667$ the post-shock conditions are imposed, whereas a reflective condition is set from $x = 0.1667$ to $x = 4$. Inflow and outflow conditions are applied at the left and right boundaries, respectively. The states at the top boundary are set to describe the exact motion of a Mach 10 shock. Figure 6 shows the pressure and density contours of the solution on a 240×60 grid. Note that compared to the results obtained by WENO-CU6-M2 [4] (their Fig. 4) a good agreement is observed. Especially, both predict a strong near-wall jet, which is usually smeared in the previous computations with the same resolution [9, 8, 6].

We then consider a shock-bubble interaction problem, when a Mach 6 shock wave in air impacts on a cylindrical helium bubble. Air and helium are treated as the same ideal gas fluid for simplicity. Numerical computations for this

problem can be found in Bagabir and Drikakis [1]. The initial conditions are

$$
\begin{cases}
(\rho = 1, u = -3, v = 0, p = 1) & \text{pre-shocked air} \\
(\rho = 5.268, u = 2.752, v = 0, p = 41.83) & \text{post-shocked air} \, , \\
(\rho = 0.138, u = -3, v = 0, p = 1) & \text{helium bubble}
\end{cases}
\tag{31}
$$

and the final time is $t = 0.15$. The computational domain for this problem is $[0, 0] \times [1, 0.5]$. Initially, the shock wave is at $x = 0.05$, and the half helium bubble of radius 0.15 is at (0,0.25). Note that a frame velocity $u = -3$ is applied to keep the bubble approximately in the center of the computational domain. Reflective conditions are applied at the lower and upper boundaries, an outflow condition is applied at the right boundary, and an inflow condition is applied to the left boundary with the post-shocked state. Figure 7 shows the pressure and density contours of the solution on a 200×100 grid. These results show a fairly good agreement with those in Bagabir and Drikakis [1] (their Fig. 6) at the same resolution. The secondary reflected shock wave and triple-wave configurations are calculated with good resolution. Note that since the WENO-CU6-M1 scheme has smaller numerical dissipation than the MUSCL scheme used in Bagabir and Drikakis [1], the present results show a less smeared bubble interface and more detailed structures near the triple-wave region.

4 Concluding remarks

In this paper we have proposed a very simple method to enforce the positivity-preserving property for general high-order conservative schemes. The method first detects critical numerical fluxes which may lead to negative density and pressure, then limits the fluxes to satisfy a sufficient condition for preserving

positivity. Though an extra time-step size condition is required to maintain the formal order of accuracy, it is less restrictive than those in previous works. In addition, since the method uses the general form of a conservative scheme, similarly as the approaches of Zhang and Shu [20], it can be applied to flows with a general equation of state and source terms in a straightforward way.

Acknowledgment

References

[1] A. Bagabir and D. Drikakis. Mach number effects on shock-bubble interaction. *Shock Waves*, 11(3):209–218, 2001.

[2] B. Einfeldt, C.D. Munz, P.L. Roe, and B. Sjogreen. On Godunov-type methods near low densities. *J. Comput. Phys.*, 92(2):273–295, 1991.

[3] J. Gressier, P. Villedieu, and J.M. Moschetta. Positivity of flux vector splitting schemes. *J. Comput. Phys.*, 155(1):199–220, 1999.

[4] X.Y. Hu and N.A. Adams. Scale separation for implicit large eddy simulation. *J. Comput. Phys.*, 2011.

[5] X.Y. Hu and B.C. Khoo. Kinetic energy fix for low internal energy flows. *J. Comput. Phys.*, 193(1):243–259, 2004.

[6] X.Y. Hu, Q. Wang, and N.A. Adams. An adaptive central-

upwind weighted essentially non-oscillatory scheme. *J. Comput. Phys.*, 229(23):8952–8965, 2010.

[7] G.-S. Jiang and C.-W. Shu. Efficient implementation of weighted ENO schemes. *J. Comput. Phys*, 126:202–228, 1996.

[8] S. Kawai and S.K. Lele. Localized artificial diffusivity scheme for discontinuity capturing on curvilinear meshes. *J. Comput. Phys.*, 227(22):9498–9526, 2008.

[9] D. Kim and J.H. Kwon. A high-order accurate hybrid scheme using a central flux scheme and a weno scheme for compressible flowfield analysis. *J. Comput. Phys.*, 210(2):554–583, 2005.

[10] R. Loubčre and M.J. Shashkov. A subcell remapping method on staggered polygonal grids for arbitrary-lagrangian-eulerian methods. *J. Comput. Phys.*, 209(1):105–138, 2005.

[11] B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for Euler equations. *Numerische Mathematik*, 73(1):119–130, 1996.

[12] L.I. Sedov. *Similarity and Dimensional Methods in Mechanics (Similarity and Dimensional Methods in Mechanics, New York*. Academic Press, 1959.

[13] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439–471, 1988.

[14] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes, II. *J. Comput. Phys.*, 83:32–78, 1989.

[15] T. Tang and K. Xu. Gas-kinetic schemes for the compressible Euler equations: positivity-preserving analysis. *Zeitschrift für Angewandte Mathematik und Physik (ZAMP)*, 50(2):258–281, 1999.

[16] E.F. Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction.* Springer Verlag, 2009.

[17] P. Woodward and P. Colella. The numerical simulation of two-dimensional fluid flow with strong shocks. *J. Comput. Phys.*, 54(1):115–173, 1984.

[18] X. Zhang and C.-W. Shu. On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. *J. Comput. Phys.*, 229:8918–8934, 2010.

[19] X. Zhang and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 467:2752–2776, 2011.

[20] X. Zhang and C.-W. Shu. Positivity-preserving high order finite difference WENO schemes for compressible Euler equations. *J. Comput. Phys.*, 231:2245–2258, 2012.
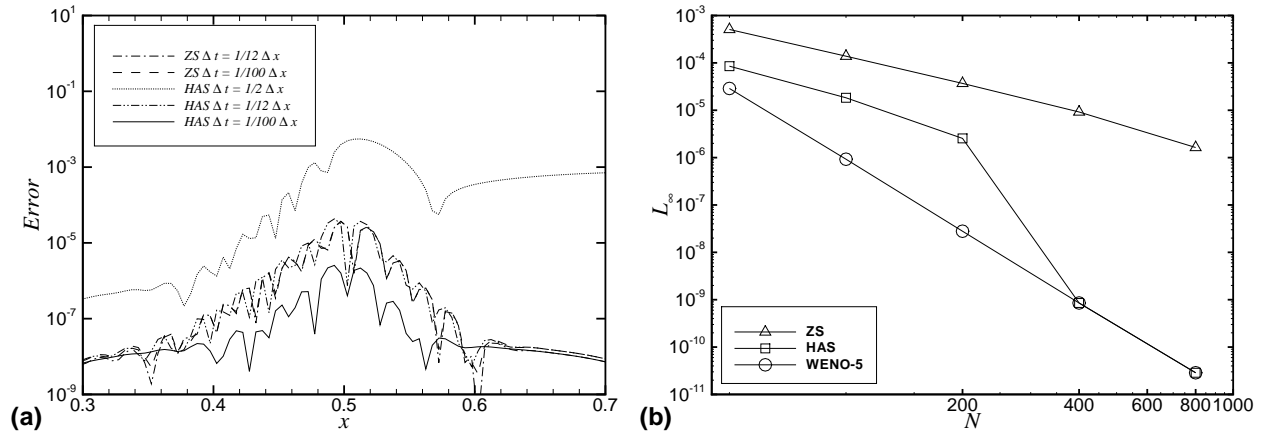
Figure 1. Linear advection problem at $t = 1$: (a) Error distribution vs. time-step sizes on 200 grid points; (b) Evolution of the L_∞ error with decreasing grid size..
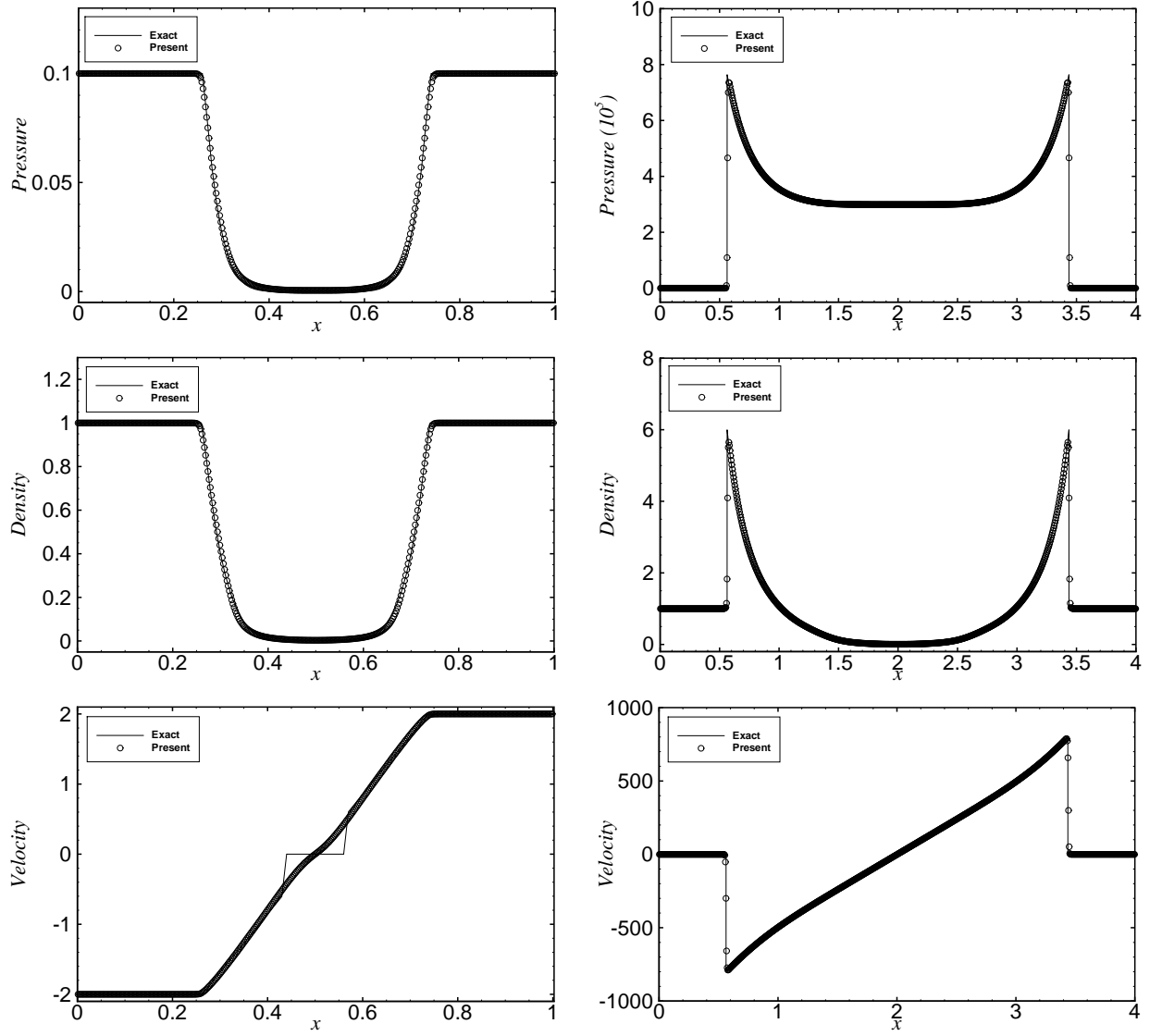
Figure 2. One-dimensional problems involving vacuum or near vacuum: (left) double rarefaction problem; (right) planar Sedov blast-wave problem.
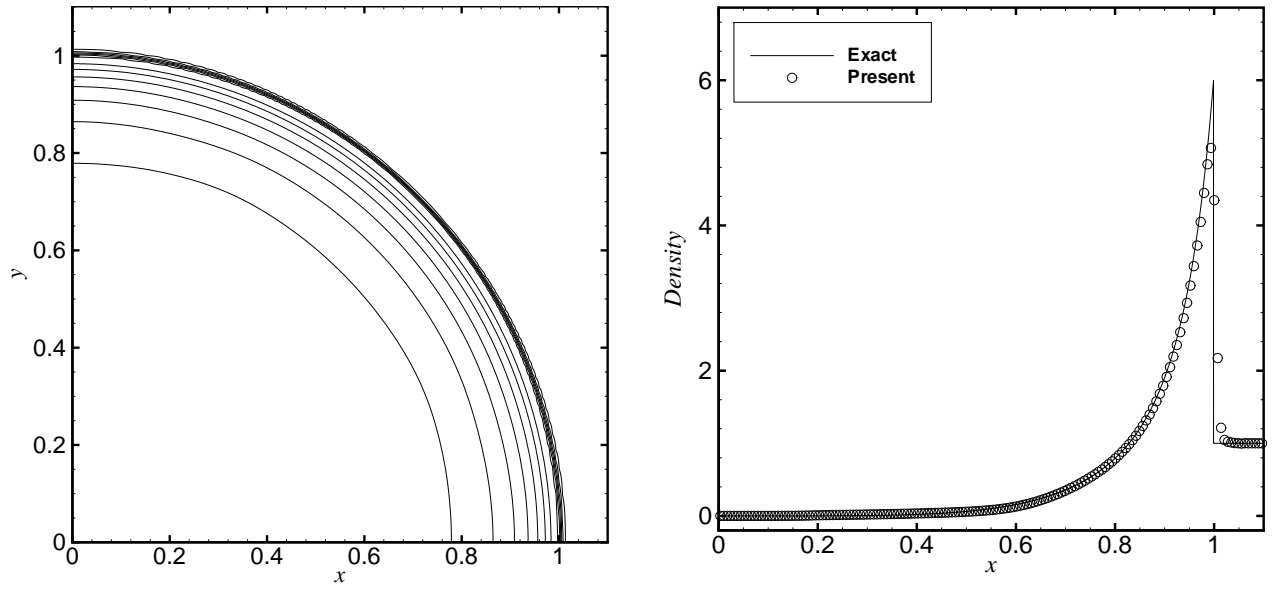
Figure 3. Two-dimensional Sedov problem: (left) 10 density contours from 0 to 6; (right) density profile along $y = 0$.
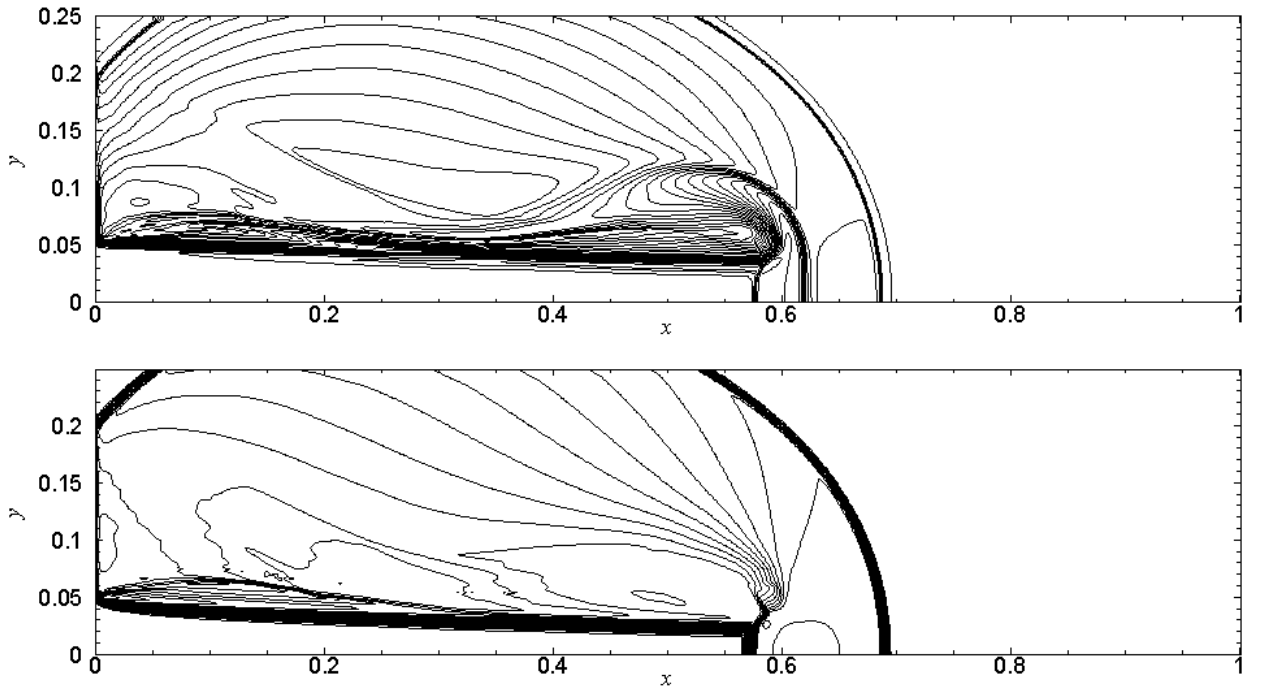

Figure 4. Mach-2000 jet problem: (upper) 30 density contours of logarithmic scale from -4 to 4; (lower) 30 pressure contours of logarithmic scale from -1 to 13.
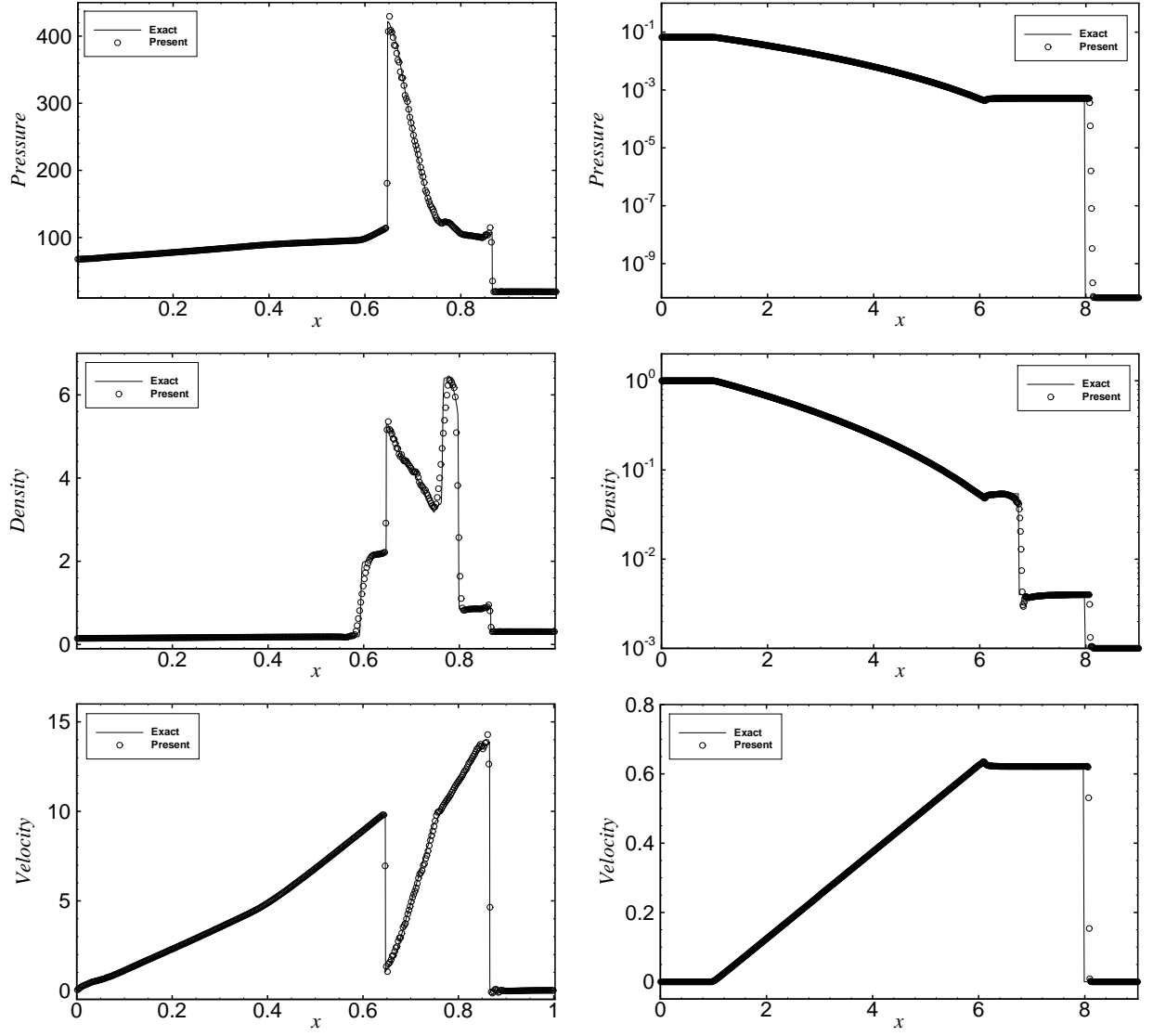
Figure 5. One-dimensional problems involving very strong discontinuities: (left) two
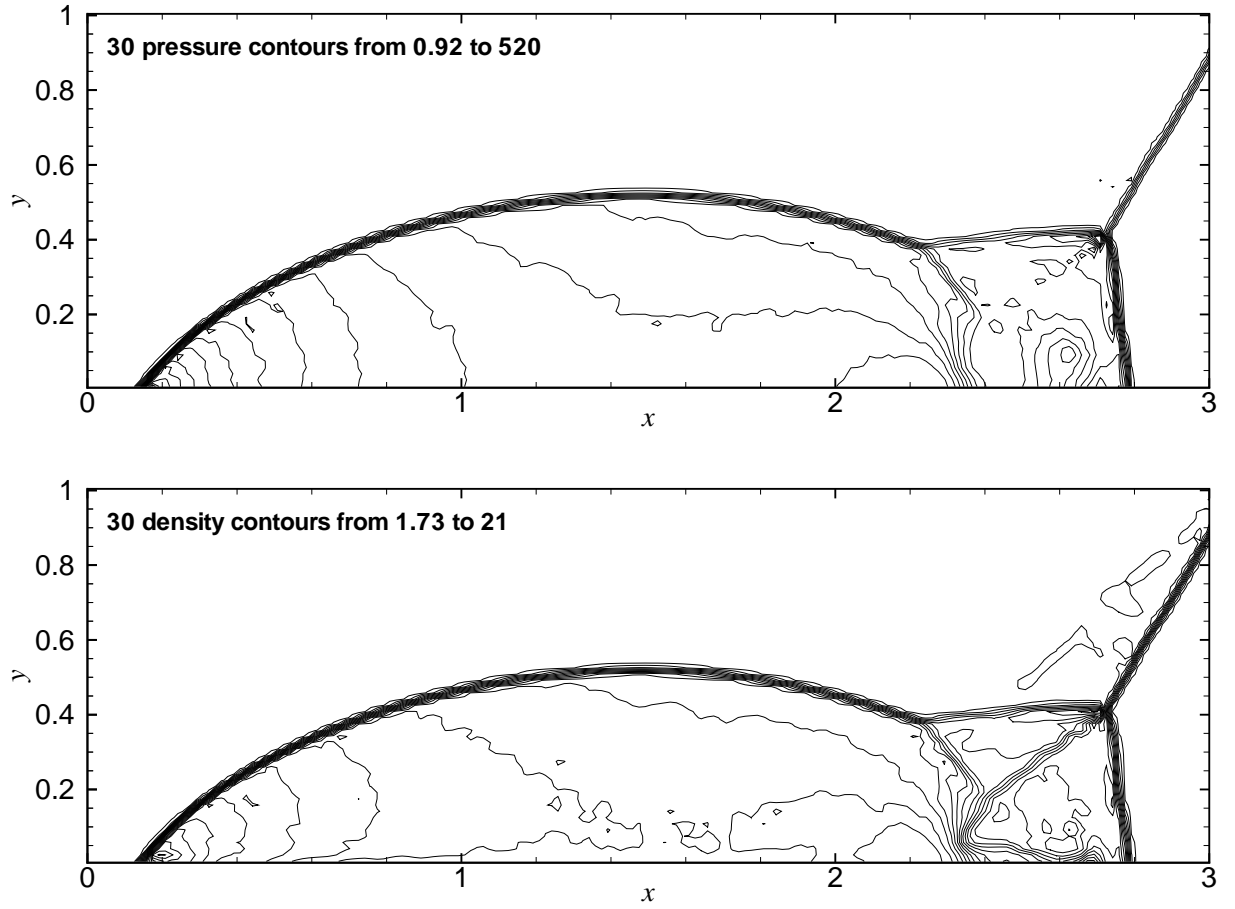blast wave problem; (right) Le Blanc shock-tube problem.

30 pressure contours from 0.92 to 520

30 density contours from 1.73 to 21

Figure 6. Double-Mach reflection of a Mach 10 shock wave at $t = 0.2$: (upper) 30 pressure contours from 0.92 to 520; (lower) 30 density contours from 1.73 to 21.
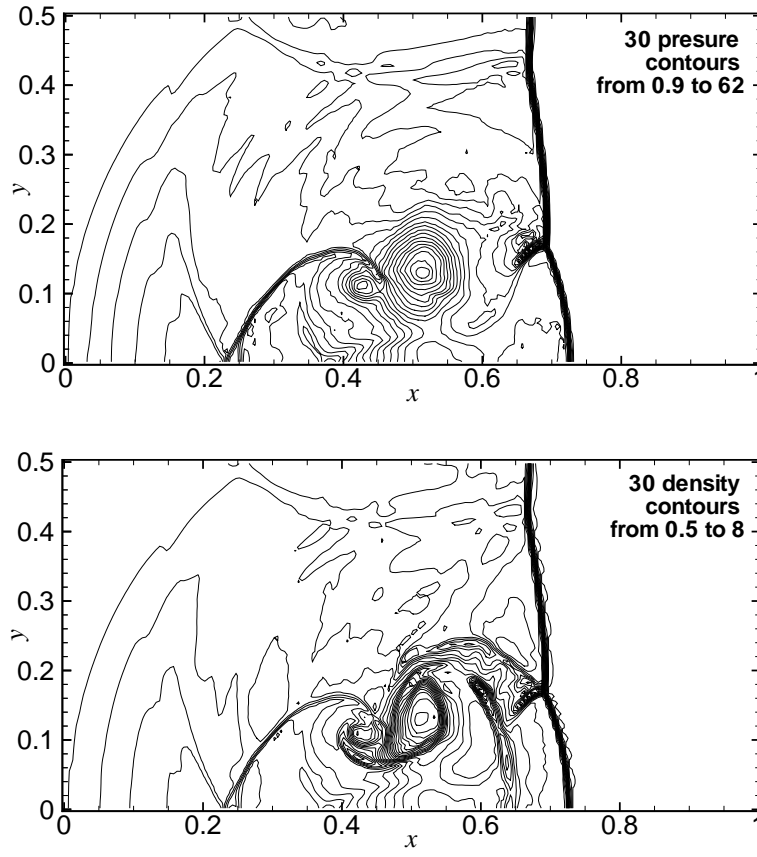
Figure 7. Shock-bubble interaction problem at $t = 0.15$: (left) 30 pressure from 0.9 to 62; (right) 30 density contours from 0.5 to 8.